09/827,432
1/26/2005

Connecting via Winsock to STN

Welcome to STN International! Enter x:x

LOGINID:ssspta1631cdl

PASSWORD:
TERMINAL (ENTER 1, 2, 3, OR ?):2

* * * * * * * * * * Welcome to STN International * * * * * * * * * *

NEWS	1		Web Page URLs for STN Seminar Schedule - N. America
NEWS	2		"Ask CAS" for self-help around the clock
NEWS	3	SEP 01	New pricing for the Save Answers for SciFinder Wizard within STN Express with Discover!
NEWS	4	OCT 28	KOREAPAT now available on STN
NEWS	5	NOV 30	PHAR reloaded with additional data
NEWS	6	DEC 01	LISA now available on STN
NEWS	7	DEC 09	12 databases to be removed from STN on December 31, 2004
NEWS	8	DEC 15	MEDLINE update schedule for December 2004
NEWS	9	DEC 17	ELCOM reloaded; updating to resume; current-awareness alerts (SDIs) affected
NEWS	10	DEC 17	COMPUAB reloaded; updating to resume; current-awareness alerts (SDIs) affected
NEWS	11	DEC 17	SOLIDSTATE reloaded; updating to resume; current-awareness alerts (SDIs) affected
NEWS	12	DEC 17	CERAB reloaded; updating to resume; current-awareness alerts (SDIs) affected
NEWS	13	DEC 17	THREE NEW FIELDS ADDED TO IFIPAT/IFIUDB/IFICDB
NEWS	14	DEC 30	EPFULL: New patent full text database to be available on STN
NEWS	15	DEC 30	CAPLUS - PATENT COVERAGE EXPANDED
NEWS	16	JAN 03	No connect-hour charges in EPFULL during January and February 2005
NEWS	17	JAN 26	CA/CAPLUS - Expanded patent coverage to include the Russian Agency for Patents and Trademarks (ROSPATENT)

NEWS EXPRESS	JANUARY 10 CURRENT WINDOWS VERSION IS V7.01a, CURRENT MACINTOSH VERSION IS V6.0c(ENG) AND V6.0Jc(JP), AND CURRENT DISCOVER FILE IS DATED 10 JANUARY 2005
NEWS HOURS	STN Operating Hours Plus Help Desk Availability
NEWS INTER	General Internet Information
NEWS LOGIN	Welcome Banner and News Items
NEWS PHONE	Direct Dial and Telecommunication Network Access to STN
NEWS WWW	CAS World Wide Web Site (general information)

Enter NEWS followed by the item number or name to see news on that specific topic.

All use of STN is subject to the provisions of the STN Customer agreement. Please note that this agreement limits use to scientific research. Use for software development or design or implementation of commercial gateways or other similar uses is prohibited and may result in loss of user privileges and other penalties.

* * * * * * * * * * * * * STN Columbus * * * * * * * * * * * * * * *

NOT ALL FILES ARE AVAILABLE AT THIS TIME. ENTER 'HELP FILE UNAVAILABLE' TO SEE THE LIST OF UNAVAILABLE FILES.

FILE 'HOME' ENTERED AT 16:55:13 ON 26 JAN 2005

=> FIL MEDLINE
COST IN U.S. DOLLARS

	SINCE FILE ENTRY	TOTAL SESSION
FULL ESTIMATED COST	1.05	1.05

FILE 'MEDLINE' ENTERED AT 16:58:09 ON 26 JAN 2005

FILE LAST UPDATED: 25 JAN 2005 (20050125/UP). FILE COVERS 1950 TO DATE.

On December 19, 2004, the 2005 MeSH terms were loaded.

Warning: The search L-number/HUMAN limit is missing from records indexed with the new 2005 MeSH (records added since December 19, 2004). Until this is corrected, include HUMANS/CT and 20041219-20051231/ED in searches to limit results to humans for this time period.

OLDMEDLINE now back to 1950.

MEDLINE thesauri in the /CN, /CT, and /MN fields incorporate the MeSH 2005 vocabulary. See http://www.nlm.nih.gov/mesh/ and http://www.nlm.nih.gov/pubs/techbull/nd03/nd03_mesh.html for a description of changes.

This file contains CAS Registry Numbers for easy and accurate substance identification.

```
=> s "trace data" and align?
         29240 "TRACE"
       1618351 "DATA"
            21 "TRACE DATA"
                 ("TRACE"(W)"DATA")
         66384 ALIGN?
L1           6 "TRACE DATA" AND ALIGN?

=> d l1 1-6 ab
```

L1 ANSWER 1 OF 6 MEDLINE on STN

AB MOTIVATION: Methylation of cytosines in DNA plays an important role in the regulation of gene expression, and the analysis of methylation patterns is fundamental for the understanding of cell differentiation, aging processes, diseases and cancer development. Such analysis has been limited, because technologies for detailed and efficient high-throughput studies have not been available. We have developed a novel quantitative methylation analysis algorithm and workflow based on direct DNA sequencing of PCR products from bisulfite-treated DNA with high-throughput sequencing machines. This technology is a prerequisite for success of the Human Epigenome Project, the first large genome-wide sequencing study for DNA methylation in many different tissues. Methylation in tissue samples which are compositions of different cells is a quantitative information represented by cytosine/thymine proportions after bisulfite conversion of unmethylated cytosines to uracil and PCR. Calculation of quantitative methylation information from base proportions represented by different dye signals in four-dye sequencing trace files needs a specific algorithm handling imbalanced and overscaled signals, incomplete conversion, quality problems and basecaller artifacts. RESULTS: The algorithm we developed has several key properties: it analyzes trace files from PCR products of bisulfite-treated DNA sequenced directly on ABI machines; it yields quantitative methylation measurements for individual cytosine positions after **alignment** with genomic reference sequences, signal normalization and estimation of effectiveness of bisulfite treatment; it works in a fully automated pipeline including data quality monitoring; it

is efficient and avoids the usual cost of multiple sequencing runs on subclones to estimate DNA methylation. The power of our new algorithm is demonstrated with data from two test systems based on mixtures with known base compositions and defined methylation. In addition, the applicability is proven by identifying CpGs that are differentially methylated in real tissue samples.

L1 ANSWER 2 OF 6 MEDLINE on STN
AB The single nucleotide polymorphism (SNP) is the difference of the DNA sequence between individuals and provides abundant information about genetic variation. Large scale discovery of high frequency SNPs is being undertaken using various methods. However, the publicly available SNP data sometimes need to be verified. If only a particular gene locus is concerned, locus-specific polymerase chain reaction amplification may be useful. Problem of this method is that the secondary peak has to be measured. We have analyzed **trace data** from conventional sequencing equipment and found an applicable rule to discern SNPs from noise. The rule is applied to multiply **aligned** sequences with a trace and the peak height of the traces are compared between samples. We have developed software that integrates this function to automatically identify SNPs. The software works accurately for high quality sequences and also can detect SNPs in low quality sequences. Further, it can determine allele frequency, display this information as a bar graph and assign corresponding nucleotide combinations. It is also designed for a person to verify and edit sequences easily on the screen. It is very useful for identifying de novo SNPs in a DNA fragment of interest.

L1 ANSWER 3 OF 6 MEDLINE on STN
AB SNPCEQer II is a graphical user interface (GUI)-based application that integrates single nucleotide polymorphism (SNP) detection, SNP analysis and SNP editing in the Microsoft Windows (R) environment. SNPCEQer II detects SNPs in DNA sequences generated by the Beckman CEQ TM 2000 XL DNA analysis system. It provides tools to analyse SNPs by inspecting and comparing **trace data** (chromatograms) around putative SNPs with that of other related DNA sequences, and it can search for those SNPs in the National Center for Biotechnology Information (NCBI) databases. SNPCEQer II can determine the mutation type of a coding SNP and generate data for submission to the dbSNP database. The SNP report can be edited and printed, as can the chromatograms. SNPCEQer II is implemented in Visual C++.

L1 ANSWER 4 OF 6 MEDLINE on STN
AB A pivotal step in electrophoresis sequencing is the conversion of the raw, continuous chromatogram data into the actual sequence of discrete nucleotides, a process referred to as basecalling. We describe a novel algorithm for basecalling implemented in the program LifeTrace. Like Phred, currently the most widely used basecalling software program, LifeTrace takes processed **trace data** as input. It was designed to be tolerant to variable peak spacing by means of an improved peak-detection algorithm that emphasizes local chromatogram information over global properties. LifeTrace is shown to generate high-quality basecalls and reliable quality scores. It proved particularly effective when applied to MegaBACE capillary sequencing machines. In a benchmark test of 8372 dye-primer MegaBACE chromatograms, LifeTrace generated 17% fewer substitution errors, 16% fewer insertion/deletion errors, and 2.4% more **aligned** bases to the finished sequence than did Phred. For two sets totaling 6624 dye-terminator chromatograms, the performance improvement was 15% fewer substitution errors, 10% fewer insertion/deletion errors, and 2.1% more **aligned** bases. The processing time required by LifeTrace is comparable to that of Phred. The predicted quality scores were in line with observed quality scores, permitting direct use for quality clipping and in silico single nucleotide polymorphism (SNP) detection. Furthermore, we introduce a new type of

quality score associated with every basecall: the gap-quality. It estimates the probability of a deletion error between the current and the following basecall. This additional quality score improves detection of single basepair deletions when used for locating potential basecalling errors during the **alignment**. We also describe a new protocol for benchmarking that we believe better discerns basecaller performance differences than methods previously published.

L1 ANSWER 5 OF 6 MEDLINE on STN

AB We present a new method for determining the consensus sequence in DNA fragment assemblies. The new method, Trace-Evidence, directly incorporates **aligned** ABI trace information into consensus calculations via our previously described representation, **Trace-Data** Classifications. The new method extracts and sums evidence indicated by the representation to determine consensus calls. Using the Trace-Evidence method results in automatically produced consensus sequences that are more accurate and less ambiguous than those produced with standard majority-voting methods. Additionally, these improvements are achieved with less coverage than required by the standard methods-using Trace-Evidence and a coverage of only three, error rates are as low as those with a coverage of over ten sequences.

L1 ANSWER 6 OF 6 MEDLINE on STN

AB A significant bottleneck in the current DNA sequencing process is the manual editing of **trace data** generated by automated DNA sequencers. This step is used to correct base calls and to associate to each base call a confidence level. The confidence levels are used in the assembly process to determine overlaps and to resolve discrepancies in determining the consensus sequence. This single step may cost as much as 4 to 8 cents per finished base. We report an approach to automated trace editing using classification trees to detect and exploit context-based patterns in trace peak heights. Local base composition and nearby peak heights account for 80% of the variations in peak heights. Classification algorithms were developed to identify 37% of automated base calls that differ from the consensus sequence. With these algorithms, 12% of the base calls had confidence levels less than 90%.